DOCUMENT RESUME

ED 039 583 CG 005 378

AUTHOR	Levine, Harold G.; Noak, John R.
TITLE	The Evaluation of Complex Educational Outcomes.
INSTITUTION	Illinois State Office of the Superintendent of Public Instruction, Springfield. Dept. of Educational Research.; Illinois Univ., Chicago. Coll. of Medicine.
SPONS AGENCY	Public Health Service (DHEW), Washington, D. C. Bureau of State Services.
PUB DATE	68
NOTE	62p.
EDRS PRICE	EDRS Price MF-$0.25 HC-$3.20
DESCRIPTORS	Educational Objectives, *Evaluation Techniques, *Measurement Techniques, Performance Tests, *Physicians, *Reliability, Simulation, *Validity

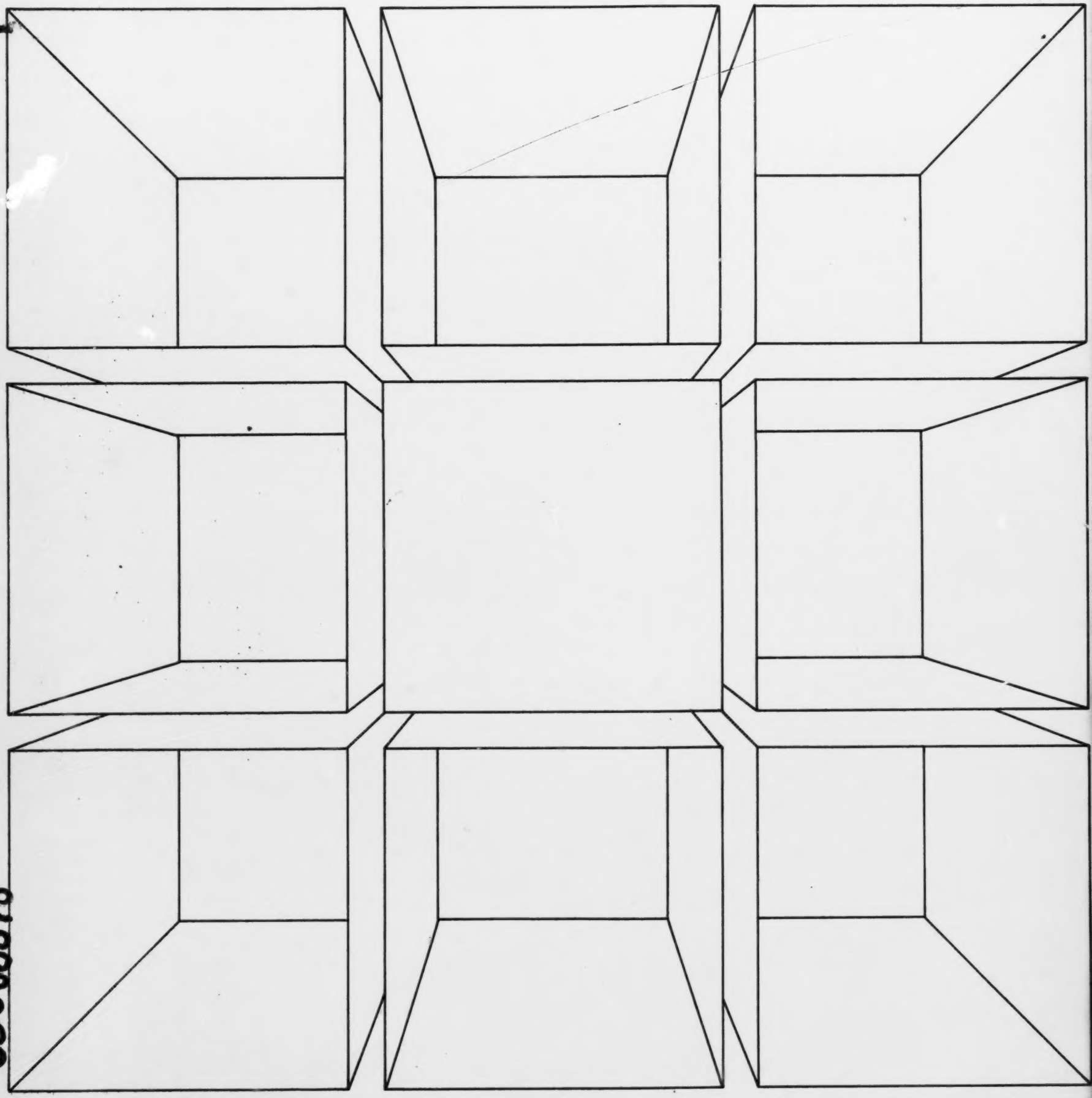
ABSTRACT

This research was designed to obtain information on the validity and reliability of three new evaluation techniques: (1) The Simulated Patient Management Problem (PMP), a written simulation exercise; (2) The Simulated Diagnostic Interview (DI), an oral exercise; and (3) The Simulated Proposed Treatment Interview (PTI), another oral exercise. These techniques are used for assessing the clinical competence of physicians. A review of related research follows a brief discussion of the purpose of the study. Following this section is a discussion on the conceptual problems in evaluating the validity of achievement tests. Chapter Five goes on to describe the three evaluation techniques listed above. This description is followed by: (1) analysis of the reliability of the techniques; (2) analysis of the construct validity of the techniques; and (3) analysis of the concurrent validity of the techniques. The last two parts of the paper present conclusions reached as a result of the study, and areas for further research. (SJ)

ED039583

The Evaluation of Complex Educational Outcomes

The Office of the
Superintendent of
Public Instruction
State of Illinois
Ray Page
Superintendent



CG005378

ED039583

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
OFFICE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL OFFICE OF EDUCATION POSITION OR POLICY.

THE EVALUATION OF COMPLEX EDUCATIONAL OUTCOMES

1968

Issued by

RAY PAGE

Superintendent of Public Instruction

Prepared by

Department of Educational Research

Division of Planning and Development

The Evaluation of Complex Educational Outcomes *

Harold G. Levine
University of Illinois
College of Medicine

John R. Noak
Department of Educational Research
Office of the Superintendent of Public Instruction

COVER DESIGN

by

Edward W. Bydalek

* A study supported in part by USPHS Grant CH 00081-01 from the Bureau of State Services (Community Health).

FOREWORD

The evaluation of complex educational outcomes remains one of the most stubborn problems facing evaluators today. Single skills, even those as multiplex as reading, can be evaluated with a fair degree of proficiency. But when native abilities and learned skills must be combined in a training program lasting four or more years, the problems of assessment increase at a geometric ratio.

Measuring the complex skills involved in being an effective teacher, physician, or scientist with any degree of accuracy is a difficult, exacting task. It is a task moreover that cannot be handled in the future as it has in the past, if evaluation is to serve any function in educational programs. Common practice has been to measure those aspects of a complex educational outcome which can be easily measured, such as recall of subject area knowledge. Equally important aspects such as problem solving, attitudes, and skills have too often either been ignored or treated spuriously.

An attempt was made, therefore, to use simulations as evaluation techniques on the theory that being closer to reality than the usual types of evaluation instruments, they would have high validity. This proved true. In the past, however, simulations have been plagued by problems of reliability. As the data illustrates, simulation techniques at the present time still have reliability coefficients lower than multiple choice examinations. But this appears to be a result of limited sampling, rather than an inherent weakness in the techniques themselves. Yet despite the reliability problem, they are able to measure proficiencies that can be measured in no other way. As the data indicates, they contribute significantly to the predictive ability of test batteries when properly used and controlled.

It is to be hoped that future models for the evaluation of complex educational outcomes using simulation techniques could be developed and implemented. A successful teacher also needs a large number of overlapping proficiencies, as do carpenters, engineers, and scientists. Unless a realistic attempt is made to stop measuring trivia and start measuring the slippery essentials of any job or profession, evaluation shall continue to be at least partially irrelevant.

Ray Page
Superintendent of Public Instruction

TABLE OF CONTENTS

Page

LIST OF TABLES . . . . . .

I. INTRODUCTION: ORIGIN OF THE STUDY 1

II. PURPOSE OF THE STUDY 3

III. REVIEW OF RELATED RESEARCH 5

IV. CONCEPTUAL PROBLEMS IN EVALUATING THE VALIDITY OF ACHIEVEMENT TESTS 9

V. DESCRIPTION OF THE TECHNIQUES 13

The Simulated Patient Management Problem--the Simulated Diagnostic Interview--the Simulated Proposed Treatment Interview--the Supervisory Rating Form--the Multiple Choice Questions--The Oral Quizzes.

VI. ANALYSIS OF THE RELIABILITY OF THE TECHNIQUES 20

Description of the Procedure--Results.

VII. ANALYSIS OF THE CONSTRUCT VALIDITY OF THE TECHNIQUES 26

VIII. ANALYSIS OF THE CONCURRENT VALIDITY OF THE TECHNIQUES 31

IX. CONCLUSIONS 41

X. AREAS FOR FURTHER RESEARCH 42

BIBLIOGRAPHY 43

APPENDIX A: CRITICAL PERFORMANCE REQUIREMENTS FOR ORTHOPAEDIC SURGEONS 45

APPENDIX B: RATING FORM 49

APPENDIX C: RESIDENT EVALUATION FORM 53

LIST OF TABLES

Table		Page
1.	Combined Rating and Sampling Reliabilities for Total Scores for Six Different Evaluation Techniques	22
2.	Reliabilities	24
3.	Mean Scores of Residents on Three Oral Examination Techniques by Level of Training	27
4.	Mean Scores of Residents on Two Written Examination Techniques by Level of Training	27
5.	Intercorrelations of Total Test Scores for Six Evaluation Techniques	30
6.	Intercorrelations of Total Test Scores with the Rating Factor Overall Competence by Year of Training	33
7.	Correlations and Multiple Correlations of Test Variables Predicting Various Rating Factors	35
8.	Percentage of Common Variance for Single Test Variables and Multiple Test Variables with Various Rating Factors	37
9.	Subtests Most Strongly Related to Selecting Rating Form Factors	39

I. INTRODUCTION: ORIGIN OF THE STUDY

About six years ago, Dr. George E. Miller, Director of the Office of Research in Medical Education, University of Illinois College of Medicine, in a routine speech to a group of physicians, challenged them on their methods of certifying the competence of physicians in specialty training. Dr. Miller told the group that if they had effective examinations, it would not be necessary to have rigid and detailed training requirements. One of the members of the audience was Dr. Charles Herndon, Chairman of the Examination Committee, American Board of Orthopaedic Surgery. He asked Dr. Miller how one might go about producing such examinations.

Eventually, in 1964, with the aid of a Public Health Service grant,[1] the Board and the Office (now called the Center for the Study of Medical Education) established a joint study of the development of competence in orthopaedics for the purpose of improving the certification procedures in orthopaedic surgery, with the hope that the findings of the study would eventually lead to increased flexibility in certifying orthopaedists.

The study required that the Board first develop a definition of competence in orthopaedics, and then develop evaluation instruments of proven validity and reliability to assess these competencies.[2] The definition of

---

[1] U. S. Department of Health, Education and Welfare, Public Health Service, Application for Research Grant, No. CH 00081-01, "The Efficient Use of Medical Manpower," Chicago, October 30, 1963, p. 8.

[2] Ibid., p. 10.

competence was developed through the critical incident technique developed by Flanagan[3] during World War II. This technique required the soliciting of incidents of effective and ineffective performance of orthopaedic surgeons from a large number of orthopaedists. These incidents were collected and categorized until no new categories emerged.[4]

The staff of the study, a team of physicians and evaluation experts, then developed some new techniques for assessing clinical competence.[5] After these techniques were developed, it became necessary to obtain some information on reliability and validity of the new techniques. The study reported here is one of a number of such studies conducted for this purpose.

---

[3]John C. Flanagan, "The Critical Incident Technique", Psychological Bulletin, 51, No. 4 (1962), pp. 327-358.

[4]J. Michael Blum and Robert Fitzpatrick, Critical Performance Requirements for Orthopaedic Surgery, Part I. Method, (Pittsburgh, Pa.: American Institutes for Research, 1965), p. 5. The 94 categories taken from Blum, pp. 8-11, are attached as Appendix A.

[5]See Christine H. McGuire and David Babbott, "Simulation Technique in the Measurement of Problem Solving Skills, "Journal of Educational Measurement, 4, No. 1 (1967), pp. 1-10, and Harold G. Levine and Christine H. McGuire, "Role Playing as an Evaluative Technique, " Journal of Educational Measurement, (In press 1968).

II. PURPOSE OF THE STUDY

The main purpose of the experiment described in this paper is to obtain information on the validity and reliability of three new evaluation techniques: The Simulated Patient Management Problem (PMP), a written simulation exercise;[6] The Simulated Diagnostic Interview (DI),[7] an oral exercise; and the Simulated Proposed Treatment Interview (PTI)[8], another oral exercise. These techniques will be described in more detail in later sections of this paper.

One question that one must always ask about a test score is its generalizability. We want to know if the individual's performance on the test can be generalized to all the situations that the test represents. This characteristic of test scores is called reliability. Reliability is defined for the purpose of this paper as the amount of variance in the measurements obtained by the test that is true variance[9]-- the extent that test scores are free of error. In this study two sources of error are especially important.

---

[6]McGuire and Babbott, "Simulation Technique," pp. 1-10.

[7]Levine and McGuire, "Role Playing."

[8]Ibid.

[9]John R. Guilford, Fundamental Statistics in Psychology and Education, 3rd ed., (New York: McGraw Hill, 1956, A, p. 436

The first is sampling error which results when an individual ranks differently on one test of the same ability than he does on another equivalent test.[10] The second, which mainly exists in tests scored by subjective judgment, is errors of rating. This study will analyze the first type of error in the written test and both types in the two oral techniques.

While reliability is important, the most important characteristic of test scores is their validity. The main purpose of a test is to provide information for the purpose of arriving at some conclusion.[11] The extent to which the scores from an instrument provide such information can be defined as the validity of the instrument.

The study will also provide information on the validity and reliability of three other techniques: multiple choice questions, oral examination quizzes, and supervisor's ratings.

---

[10] Ibid., pp. 444-445.

[11] E. F. Lindquist, ed. Educational Measurement, chap. 14, "Validity" by Edward E. Cureton, (Washington, D. C.: American Council on Education, 1951), p. 622.

III. REVIEW OF RELATED RESEARCH

The three techniques studied are essentially attempts to gather information on competency by simulating certain aspects of a physician's work. Work sample tests are probably as old as work. It is probable that employers have asked carpenters to nail a few boards together before hiring them, and bank tellers to add up columns of figures for hundreds of years. The first systematic, scientific attempts to predict performance in complicated professions were made by German and British military psychologists.[12] Most of these tests were not adequately validated because it was difficult to obtain any meaningful estimate of effectiveness of job performance for men in the military.[13] After the war, the British Civil Service used a three-day house party to assess candidates for high positions in the service. Correlations between final assessments at the end of the house party and ratings of job performance were .50 - .65. This is quite high for such assessments. In a two-year follow-up using supervisor's ratings as criteria, the correlations between such ratings and written abilities tests had a median of about .12. The median of the correlations of performance tests and interviews with the criteria was .41.[14] Cronbach states, 'Evidently, the impressionistic procedure identified aptitudes the paper-and-pencil tests did not."[15]

---

[12]Lee J. Cronbach, Essentials of Psychological Testing, (New York, Harper Bros., 1960), p. 567.

[13]Ibid.

[14]Ibid., p. 583.

[15]Ibid.

These performance tests were for the most part job replicas of Civil Service paperwork, committee tasks, and group discussions. It is important to note that those evaluating the examinees had a clear and agreed upon idea of what they were looking for based upon a thorough analysis of the positions involved.[16]

The only thorough analysis of attempts to evaluate physicians in terms of evaluation of characteristics required for effective job performance was done by Holt and Luborsky[17] on psychiatrists in training (residents) at the Menninger Clinic in Kansas. In this study, the raters tried to predict the effectiveness of residents at the end of their training on the basis of tests and interviews with the residents entering training which probed certain psychological traits developed on the basis of Freudian personality theories. The criteria used were supervisor and peer ratings of job performance. The results of this study were unsatisfactory as the average correlation with job performance for combined information from tests was .27 and from interviews was .24.[18]

The present study differs considerably from either the studies of the British Civil Service type orthe studies of the Menninger Clinic type. Those studies attempted to isolate some characteristics or traits of the individuals involved which were prerequisites for effective job performance, and to predict job success on the basis of performance on these traits.

---

[16] Ibid.

[17] Robert R. Holt and Lester Luborsky, Personality Patterns of Psychiatrists (New York: Basic Books, 1958), cited by Cronbach, Essentials, p. 584.

[18] Cronbach, Essentials, p. 584.

Such studies are hampered by the fact that the relationship between the trait analyzed and job performance must be assumed, and these assumptions may be false. For example, the psychiatrists assumed that such psychological traits as anxiety, etc., would bar a man from performing successfully as a psychiatrist. This turned out not to be true. Often individuals are able to overcome their weaknesses in performing a task. As Cronbach points out, people often behave according to the requirements of a job even though the personal predilections may be against such behavior. A man who habitually slouches can still learn how to stand straight if his career as an army officer requires it.[19]

For this reason, it is probable that job replica tests which require the examinee to play roles very similar to those required on the job would be more successful than results based on psychological traits. This probably explains the success of the British Civil Service tests. It is comparatively easy to develop job replica tests for civil servants before they have received training because the British educational system trains people to perform similar tasks before entry into the career service. It is quite difficult to do this for selecting psychiatric residents since they need to learn a great deal about their jobs before they can perform any tasks which have much similarity to the duties of trained psychiatrists.

---

[19] Ibid., p. 585.

The techniques developed for the American Board of Orthopaedic Surgery are being developed for certification rather than selection. It is possible then to use criteria of present job performance to validate the techniques rather than future performance. This is a study of concurrent validity rather than predictive validity.[20]

Since the physicians are already trained, it is much easier to devise work sample tests based upon a detailed analysis of the critical performance requirements of the position than it would be to devise such tests for untrained persons to use for selection purposes. The high validities of work sample tests achieved in the British Civil Service suggests that such an approach may be useful in the certification of physicians.

---

[20]See Cronbach, Essentials, pp. 108-109, for a detailed discussion of these terms.

IV. CONCEPTUAL PROBLEMS IN EVALUATING THE VALIDITY OF ACHIEVEMENT TESTS

It is easy to understand why it is necessary to evaluate selection instruments such as those devised by the psychiatrists in the Menninger Clinic study. Unless one checks up on the results of the tests, no one is certain that such traits as introversion or compulsiveness have any relationship to ability to function as a psychiatrist. But why should it be necessary to check on an achievement test? It is true that a great many achievement tests can be validated on the basis of content alone. If a teacher had as her goal that the pupils should be able to add or subtract, then few would quarrel with a paper and pencil test which required the pupils to add or subtract as long as it sampled most of the possible number combinations pupils must encounter. Unfortunately, a great many teacher's goals are not susceptible to being directly sampled by means of a test. Her goal may be the ability to use arithmetic effectively in everyday life. Pupils would be expected to be able to balance a checkbook or restaurant check correctly. The teacher may not be sure that pupils who can add or subtract on a simple mathematics test really could perform similar tasks in real life. If some of her pupils were waitresses (we can assume it was an EMH[21] class), and she collects restaurant checks from customers--difficult as that may be--she might still be in the dark about her pupils' true abilities. The sample of checks may be too small or the pupils may have

---

[21] Educationally Mentally Handicapped

been assisted by customers and fellow employees. If, however, the restaurant checks agreed substantially with the test of addition and subtraction, then the teacher could be reasonably certain that both were reasonable estimates of the pupils' mathematical abilities.

This example illustrates the difficulty of any statistical validation of evaluation instruments. The measurements used as a criterion are as subject to lack of reliability and validity as are the scores from the instrument being validated.

If these problems exist for such a simple test as fourth grade addition and subtraction, one must realize how much more serious they are for tests in the field of medicine. For example, the PMP's require the examinee to make judgments about a simulated patient in the diagnosis and treatment of a patient's complaints. It is assumed that the results of such exercises have some relationship to how the examinee would actually treat patients. The examinee taking a PMP, however, undertakes the task under conditions quite different from the conditions he encounters when he faces a patient. How important are these differences? It is difficult to answer this question. The approach used in this study is to obtain ratings from supervisors in a position to observe the habitual performance of examinees to compare their ratings with their scores on the PMP's. The main difficulty with this technique of concurrent validation is that the ratings may contain as much error as the simulation exercises.

There are at least five sources of error.

(1) The supervisors may disagree with one another or even with themselves on the standards they should employ.

(2) The supervisors may not have the opportunity to observe a particular component of behavior at all. For example, few supervisors of training programs ever observe a resident interviewing patients.

(3) The supervisors may not have observed the resident at a particular task for a sufficient number of times to generalize about the behaviors discussed.

(4) The supervisors may misunderstand or misinterpret the instructions on the rating form.

(5) The supervisors may tend to rate the person who particularly impressed or failed to impress them on one trait high or low on all traits (halo effect).

Because of these problems, it is difficult to treat either the test or the ratings as definitive estimates of the examinee's abilities. However, it is hoped that if one explores the relationship between the two types of evaluation procedure, it is possible to obtain valuable information on the aspects of performance that both are measuring.

Since concurrent validation presents such problems, psychometricians have devised other techniques to obtain information on the validity of the data provided by evaluation instruments. One technique which can be used is to develop a hypothesis about the nature of the concepts being evaluated by the test and then design an experiment to test the hypothesis. Such a

procedure is called construct validation.[22] An example of such a procedure would be to administer a test designed to be a measure of surgical skill to a mixed group of surgeons and interns. One would expect that if the test were valid, the surgeons would do better than the interns.

---

[22]Cronbach, Essentials, pp. 104-105.

V. DESCRIPTION OF THE TECHNIQUES

The 94 categories of the critical performance requirements for orthopaedic surgeons which were developed through the critical incident study mentioned earlier were classified into the following nine large categories:

I. Skill in Gathering Clinical Information
II. Effectiveness in Using Special Diagnostic Methods
III. Competence in Developing a Diagnosis
IV. Judgment in Deciding on Appropriate Care
V. Judgment and Skill in Implementing Treatment
VI. Effectiveness in Treating Emergency Patients
VII. Competence in Providing Continuing Care
VIII. Effectiveness of Physician-Patient Relationship
IX. Accepting the Responsibilities of a Physician[23]

Each of the three experimental techniques were specifically designed to evaluate one or more of the categories above.

The Simulated Patient Management Problem (PMP)

This technique consists of two booklets. In one booklet is listed a statement of a problem and a number of alternate procedures. The other booklet contains an answer sheet covered with an opaque overlay. The examinee is first required to read the case description which is usually

---

[23]Blum and Fitzpatrick, Critical, pp. 8-11.

very brief. A typical description might be, "A 55-year-old man comes to you complaining of pain in his back." The examinee's next task is to make a decision on several possible procedures. The instructions might read:

You would NOW (Select ONLY ONE):

1- Take a history
2- Administer a physical examination
3- Order laboratory tests
4- Admit the patient to the hospital

If the examinee decides to take a history, he would erase the overlay from the section of the answer sheet opposite the number 1. The answer sheet would state: "Go to Section A". In Section A of the test booklet would be listed a number of questions such as, "Where does it hurt?" At the end of the history section the examinee would again be confronted with decisions as to his next course of action. The examinee works through the problem until he either kills the patient, cures him, or loses him to another physician.

As can be seen, this technique is specifically designed to provide information on I. Skill in Gathering Clinical Information, III. Competence in Developing a Diagnosis, and VI. Judgment in Deciding on Appropriate Care.

The technique is scored by giving a weight to each erasure according to a scale derived by a criterion group of physicians. Those decisions which are regarded as definitely beneficial to the health of the patient are given positive weights. Those decisions which are regarded as definitely detrimental to the health of the patient are given negative weights. All other decisions are given zero weights. Each PMP therefore yields three scores:

(1) The sum of the positive weights--positive score.

(2) The sum of the negative weights--negative score.

(3) The sum of positive and negative scores--net score.[24]

The Simulated Diagnostic Interview (DI)

This technique is, in effect, an oral version of the PMP. The examinee is given a brief description of a patient's complaint. He then plays the role of a physician and elicits a history of the complaint from the examiner who plays the role of a patient and who has memorized the details of a case. After the history is completed, the examinee may request other diagnostic information or the physical examination and laboratory data. At the end of 12 minutes, the examinee is requested to stop and is given 3 minutes to present his diagnostic impressions.[25]

This technique is designed to provide information on I. Ability to Gather Information, III. Competence in Developing a Diagnosis, and to a lesser extent on VIII. Effectiveness of Physician-Patient Relationship.

The technique is scored on an impressionistic basis in terms of five factors. Each factor is described in some detail.[26] The rater uses a 12 point scale with 1-3 poor, 4-6 adequate, 7-9 good, and 10-12 excellent.

---

[24]McGuire and Babbott, "Simulation Technique," pp. 1-10.

[25]Levine and McGuire, "Role Playing."

[26]The Rating Form is attached to this paper as Appendix B.

The five factors are:

(1) Ability to gather pertinent information.

(2) Ability to communicate with patients.

(3) Efficiency at gathering information.

(4) Ability to arrive at a diagnosis.

(5) Overall competence.

The Simulated Proposed Treatment Interview (PTI)

In this technique the examinee plays the role of a physician and the examiner that of a patient. The examinee has three minutes to thoroughly familiarize himself with the details of a case, and then it is his task to explain the treatment outline in the case description to the "patient".[27] This technique is mainly concerned with evaluating VIII. Effectiveness of Physician-Patient Relationship. The rating system is the same as that used for the Diagnostic Interview.[28]

The factors are:

(1) Effectiveness of the examinee's statements.

(2) Effectiveness of the examinee's manner.

(3) Effectiveness of the interaction between patient and examinee.

(4) Overall competence.

---

[27] Levine and McGuire, "Role Playing."

[28] The Rating Form is attached to this paper as Appendix B.

The Supervisory Rating Form

This form was developed to evaluate the traits required to perform adequately on most of the components of competence listed in the critical incident study. It consists of brief descriptions[29] of the following factors:

I. Ability to recall factual information concerning general medicine and orthopaedic surgery.

II. Ability to use information to solve problems.

III. Ability to gather clinical information.

IV. Judgment in deciding on appropriate treatment and care.

V. Skill in surgical procedures.

VI Relating effectively to patients.

VII. Relating effectively to colleagues and other medical personnel.

VIII. Demonstrating the moral and ethical standards of a physician.

IX. Overall competence as a physician.

The form was also rated on a 12 point scale.

The Multiple Choice Questions

The Board and the American Academy of Orthopaedic Surgeons which administers an examination to all orthopaedic residents, have been using a traditional four-or-five-option, single answer multiple choice examination as the main examination for the past few years. In 1964, a team of orthopaedic

---

[29] A copy of the form is attached as Appendix C.

surgeons analyzed the written examinations (which were for the most part multiple-choice questions) in order to determine the type of mental processes demanded of the examinees by the test questions. The results were summarized as follows:

(1) Over half the questions were classified as recall by all the experts.

(2) Less than 25% of the questions were thought by any expert to involve even simple interpretation of data, application of principles, or evaluation.[30]

The Oral Quizzes

The Board has given five, one-half hour oral quizzes to the candidates for certification for many years. The oral examinations, which are administered by large numbers of practicing orthopaedists, have always been considered the heart of the examination--the portion which really forced the candidates to demonstrate their competence.

The planning and administration of the examinations are very loose. Examiners are simply invited to come and examine the candidates in a particular subject matter area with the assistance of another examiner.

---

[30]George E. Miller, Christine H. McGuire, and Carroll B. Larson, "The Orthopaedic Training Study--A Progress Report, "Bulletin of the American Academy of Orthopaedic Surgeons, 13 (1965), pp. 8-11.

These examinations were also subjected to a process similar to that conducted on the written examinations. The findings were summarized as follows:

(1) Nearly 70% of the questions asked required only recall and recognition of isolated fragments of information.

(2) Fewer than 20% of the questions asked required demonstration of interpretive skill.

(3) Only 13% of the questions included any element of problem solving.[31]

The findings on both traditional techniques seem to indicate that they assess mainly the store of information required to perform effectively in the areas of competence outlined by the critical incident study. The research on the validity of the experimental techniques were predicated on the assumption that although such assessment was valuable, it did not go far enough. The experiment discussed below was conducted to test this assumption.

---

[31]Miller, McGuire, and Larson, "The Orthopaedic Training," p. 9.

VI. ANALYSIS OF THE RELIABILITY OF THE TECHNIQUES

Description of the Procedure

The data reported on the validity and reliability of the techniques in this report are based upon the examination given to residents at all levels of training in November, 1966. This examination consisted of a multiple-choice section and a patient management problem section, and was taken by 1,529 residents, approximately 90% of all the residents in orthopaedics in the United States. In order to gain estimates of the statistical quality of the other techniques, special arrangements were made to administer oral examinations to a subset of this group.

The residents in 23 training programs located in five areas of the country; Rochester, Minnesota, New York City, San Francisco, Boston, and Chicago; were canvassed and asked to participate in a study of oral examination techniques.

Of the approximately 500 who were eligible to participate, 233 agreed and took the oral examination. This is obviously not a random sample, and, therefore, the data cannot be generalized beyond those 233 without considerable caution. On the other hand, the data on multiple-choice and PMP scores of the 233 does not differ markedly from scores for the population of residents. All the 233 candidates took an hour examination. One half-hour consisted of a traditional oral examination quiz in adult orthopaedics administered in the usual fashion followed by the Board but using only one examiner.

The other half-hour consisted of the Diagnostic Interview and Proposed Treatment Interview administered together by one examiner. In addition, small groups of approximately 30 candidates were observed by two examiners to provide a measure of rating reliability. Two other subsets of 25 residents either took two Simulated Interviews as well as one Adult Oral Quiz, or two Adult Oral Quizzes and one administration of the Simulated Interview. These subsets provided estimates of the combined effects of sampling and rating reliability.

The reliability of the ratings was estimated by sending two forms to the supervisors of the training programs for each of the 233 residents who took the oral examinations. The request was made that two individuals who were in a position to know the resident would fill out the forms independently. Most of the programs fulfilled this request. Correlation of the two questionnaires provides estimates of the combined effects of rating and sampling.[32]

Results

The results of the analysis of reliability are summarized in Table 1. Some notes on the data in Table 1 follow.

(1) The reliability of the multiple choice questions were computed by the Kuder-Richardson 21 method.[33]

---

[32]The extent to which such a correlation reflects sampling reliability depends upon the extent that the two raters observed different incidents in the performance of the residents.

[33]Guilford, Fundamentals, p. 455.

(2) The reliability of the PMP's was computed by Angoff 12 formula[34] which is an estimate of the internal consistency of the problems. This formula estimates the relationship between problems used in the 1966 In-Training Examination and another set of problems using similar content.[35]

TABLE 1.-- Combined Rating and Sampling Reliabilities for Total Scores for Six Different Evaluation Techniques

Evaluation Technique	N	Reliability
Multiple Choice	1,529	.90
PMP Total	1,529	.90
Diagnostic Interview	25	.14
Proposed Treatment Interview	25	.49
Adult Oral Quiz	25	.54
Ratings	190	.73

This formula means essentially the same thing as the Kuder-Richardson formula cited above. However, the PMP's are quite different from multiple choice tests and, therefore, the relationship between the PMP's and ratings is probably more limited. The ratings and multiple choice questions are both based on a number of independent sources of

[34] William H. Angoff, "Test Reliability and Test Length, "Psychometrika, 19 (1953), pp. 1-16.

[35] Arieh Lewy and Christine McGuire, "A Study of Alternative Approaches in Estimating the Reliability of Unconventional Tests," speech read at the Annual Meeting of the American Educational Research Association, Chicago, February 18, 1966, p. 11.

information. The PMP's are based on only a few. For example, the ratings may deal with the observer's impressions of the candidate's ability to handle dozens of illnesses. The PMP reports reliably on his ability to handle one or two illnesses. The correlation between PMP's dealing with two diseases and another set of PMP's dealing with two different diseases is probably much lower than .90.[36] This is also true of multiple choice tests, since a multiple choice test on children's orthopaedics will not correlate very high with multiple-choice tests on adult orthopaedics. However, the multiple-choice total contains all the important subject areas while the PMP's total cannot.

(3) The reliability of the ratings was computed in Table 1 by correlating the ratings and then correcting the correlations by use of the Spearman-Brown formula.[37] This was done because the combined ratings were used as criteria in the section on concurrent validity in this paper.

(4) Both the rating and combined rating and sampling reliabilities were computed although only the estimate showing the combined effect of both errors is shown in Table 1. Table 2 below gives the complete data for the three orals. The data in Table 2 are somewhat suspect because of the small size of the samples. It is heartening, however, that analysis of the

---

[36] Ibid., pp. 13-14.

[37] Guilford, Fundamentals, p. 454.

TABLE 2.--Reliabilities

Test	Duration	N	Reliability of Rating *	N	Reliability of Rating and Sampling Combined **
Diagnostic Interview	15 min.	33	.64	25	.14
Proposed Treatment Interview	10 min.	33	.55	25	.49
Adult Oral Quiz	30 min.	30	.72	25	.54

*Computed by having the same test observed by two examiners and correlating results.

**Computed by having the examinee take tests with different content from different examiners and in relating the results.

rater reliability of the simulated interviews conducted on the 1966 Orthopaedic Certifying Examination produced very similar results.[38]

It is interesting to note the strong effect that case differences have on the Diagnostic Interview. This effect may result from the fact that first and second year residents had large gaps in their knowledge. The sampling reliability for the Diagnostic Interview is probably higher for candidates for certification. In any case the Board has decided to use a number of different cases in arriving at scores on its certification examination.[39] The high reliability of the 15-minute Proposed Treatment Interview is very

---

[38]Levine and McGuire, "Role Playing."

[39]Charles F. Gregory, Letter to Examination Committee, American Board of Orthopaedic Surgery, September, 1967.

heartening. In this case the nature of the treatment to be explained appears not to be important. The Adult Oral is somewhat affected by the cases used but not as seriously as the Diagnostic Interview. The fact that the test is twice as long as the DI and contains more different cases probably explains the reason for these results.

In any case, these figures are too low to allow these tests to be used independently to assess individuals, but they are high enough to suggest the use of these tests as part of batteries of tests for certification purposes.[40] This is the way the Board intends to use the Simulated Interviews.[41]

One further note on the data in Table 2. The reliabilities given are direct correlations between the two raters and are thus estimates based on the reliability of the ratings of one rater. If two raters are used and the scores pooled, the reliabilities, of course, will be higher.

---

[40] Guilford, Fundamentals, p. 473.

[41] Gregory, Letter.

VII. ANALYSIS OF THE CONSTRUCT VALIDITY OF THE TECHNIQUES

To the extent that the instruments are measuring some abilities which are related to the objectives of the training programs, then those with more training should perform better than those with less training. The analysis of the 1966 In-Training Examination presented an excellent opportunity to explore this aspect of validity since the examinees were at different levels of training. Table 3 summarizes the results of analyzing the data on the three oral examinations by level of training.

All of these tests show growth through the training period. It is not surprising that the PTI shows the least growth. Most analysts of training programs would concede that there is little formal effort to improve the resident's skills in the area of competence sampled by this technique.

The data on the multiple-choice and PMP techniques analyzed in a similar fashion are presented in Table 4. Note that the breakdown on the PMP subtests has also been included.

The data in Table 4 were not subjected to tests to statistical significance, but in view of the large N's, the differences in the multiple-choice data would certainly be found significant. The PMP data obviously would not.

TABLE 3.--Mean Scores of Residents on Three Oral Examination Techniques by Level of Training

Level of Training	N	Diagnostic Interview		Proposed Treatment Interview		Adult Oral Quiz	
		Mean	SD	Mean	SD	Mean	SD
1st year	29	5.4	2.2	6.2	2.8	65%	9%
2nd year	75	6.8	1.7	6.9	2.8	70%	13%
3rd year	50	6.9	2.5	6.5	2.8	75%	12%
4th year	79	7.6	2.5	7.5	2.6	80%	10%
Total	233	6.9	2.8	6.9	2.6	74%	12%

* These tests were scored on a 12 point scale with 1-3 poor, 4-6 adequate, 7-9 good, 10-12 excellent.

NOTE: A multivariate analysis was run on all of the subjects of both simulated interviews and the adult oral examination. This analysis showed differences at the .0001 level. Univariate analysis showed differences significant at the .001 level for both the Diagnostic Interview and the Adult Oral. The Proposed Treatment Interview barely missed significance (P = .08).

TABLE 4.--Mean Scores of Residents on Two Written Examination Techniques by Level of Training

Techniques	1st Year = 256	2nd Year = 531	3rd Year = 345	4th Year = 390
N				
Multiple Choice Total	48%	52%	57%	61%
PMP Total	24%	25%	22%	23%
Problem I Diagnostic Net	62%	63%	55%	59%
Problem II Therapy Net	-10%	-7%	-7%	-4%
Problem II Diagnosis Net	7%	7%	6%	5%
Problem III Therapy Net	17%	20%	20%	15%

The question arises as to why the PMP technique does not show improvement during training. Studies conducted on the 1965 Orthopaedic In-Training Examination[42] and the 1966 Orthopaedic Certifying Examination[43] indicate that problems dealing with diagnosis do not discriminate between levels of training, but those that deal with treatment do. This pattern seems to hold true for Problem I, but not for Problem II. The discrepancy in Problem II probably results from the nature of the problem in which choice of treatment depends so heavily on diagnosis that the poor diagnostic scores achieved on the problem made it possible to demonstrate therapeutic judgment.

It is not difficult to understand why this discrepancy should exist; the main emphasis in the scoring of diagnostic type problems is on thoroughness. The criterion group gives positive weights to a number of diagnostic procedures which are needed to rule out diagnoses which may be less common than the ones usually associated with a syndrome, but still common enough to affect a significant number of patients.

The resident, however, usually works in a charity hospital in which the emphasis is on discrimination rather than thoroughness. The accolades

---

[42]Harold G. Levine, "Analysis of the Construct Validity of Two Simulation Techniques, " (Chicago: Center for the Study of Medical Education, University of Illinois College of Medicine, 1967), p. 12 (Dittoed report).

[43]Harold G. Levine, "Report on the January 1966 Orthopaedic Certification Examination, " (Chicago: Center for the Study of Medical Education, University of Illinois College of Medicine, 1967), p. 67 (Dittoed report).

go to the resident who makes quick diagnoses and saves time and money while doing so. Furthermore, under the systems of training used by most training programs he rarely has a chance to follow up his cases and see the consequences of his failures.[44]

Another approach to the construct validation of the examinations is to explore the correlations among them. Those techniques which content analysis would indicate were measuring different aspects of competence should correlate low with each other. Table 5 presents the intercorrelations of all six techniques investigated in this study.

Note the PTI correlates quite high with DI. This fact probably results from administering both tests in the same half-hour period, using the same examiner. When two different examiners are used, the correlation between the PTI of one and the DI of the other is very low. In recognition of the effect of including the two techniques designed to measure different things in the same examination period, the Board has changed the method of administering these two techniques.[45]

---

[44]These remarks are based on a number of talks with orthopaedists in both practice and university settings. Especially helpful have been Dr. Brian Huncke, a practicing orthopaedist, who gives a day a week of his time to the Department of Orthopaedic Surgery, University of Illinois College of Medicine and the Center for the Study of Medical Education, and Dr. Floyd H. Bliven, Chairman, Department of Orthopaedic Surgery, Medical College of Georgia.

[45]Gregory, Letter.

It is also interesting to note the high correlation between the Multiple-Choice test and the Adult Oral. This is not surprising in view of the process analyses discussed earlier.[46]

The other correlations are low as would be expected from their content analyses and the low reliabilities of some techniques, especially the Diagnostic Interview. It is interesting to note, however, that all the techniques have spectacularly low relationships with the ratings. Does this mean that the scores on these techniques are not related to observations of habitual performance? This question involves concurrent validity of the techniques and is properly the subject of the next section.

TABLE 5.--Intercorrelations of Total Test Scores for Six Evaluation Techniques

	PTI	Adult Oral	Multiple Choice	PMP	Rating Factors
Diagnostic Interview	.51**	.18	.26**	.06	.10
Proposed Treatment Interview	---	.20	.27**	-.04	.17
Adult Oral		---	.44*	-.09	.22*
Multiple Choice				.01	.26**
PMP					-.01

* Significant at .05 level
** Significant at .01 level

[46] Miller, McGuire, and Larson, "The Orthopaedic," p. 9.

VIII. ANALYSIS OF CONCURRENT VALIDITY OF THE TECHNIQUES

It would not be surprising if the ratings failed to have any significant relationships to the tests. As was pointed out earlier, there are many reasons why they may not correlate with test scores. There is, however, one other reason which does not require that either the ratings or the test scores be invalid. Perhaps the best way to explain this would be to use a rather elaborate analogy--with apologies to the reader who lacks familiarity with baseball.

Assume that a group of sports writers were polled and asked to list the greatest baseball players of the last 40 years. At the same time, someone digs through the records and obtains such data as batting averages, fielding averages, runs batted in, etc. Assume further that they find (as they probably would) that no one of these correlated very highly with overall estimates of greatness as a baseball player.

There would be several reasons for this:

(1) The overall competence of a baseball player depends on the combination of a number of rather divergent skills; ability to hit often, ability to hit far, ability to hit with men on base, ability to field, etc. The statistics, batting averages, runs batted in, etc., deal with only one of these abilities. It is not surprising that any one statistic would not relate highly with overall competence.

(2) It would be improper to mix older ballplayers with younger ones because style of play has changed through the years. Raters take this into account in estimating greatness, but the statistics cannot unless special arrangements are made to do so. For example, most sports writers rate Ty Cobb and Babe Ruth as equally great, yet Ty Cobb hit few home runs and Babe Ruth struck out a great deal. Criteria based upon home runs and lack of strikeouts are unfair to one or the other.

(3) The overall competence score is too broad to correlate very highly with any criteria. If the sports writers were selecting the best pitcher and the best hitters, data which correlated low with overall competence would correlate quite well with these criteria.

(4) Factors exist, such as leadership ability, which would never correlate with any of the statistical data. The existence of such factors would naturally lower the possible relationships between statistical data and sports writers' ratings.

This analogy suggests several means of analyzing the relationship between the scores on the evaluation techniques and supervisory ratings. One step is to separate the residents by year of training. This should be done <u>first</u>, because the supervisors may use different criteria to evaluate residents at various levels of training and <u>second</u>, because the supervisors

have had an opportunity to observe the third and fourth year residents for a longer time, and the ratings for these residents should be more reliable. Table 6 presents the correlations when the residents are separated in this fashion.

It is interesting to note how all of the correlations improve for the third and fourth year residents. This may reflect the two reasons listed above or a third reason that may exist. It may be that some tests such as the Diagnostic Interview are not appropriate for residents early in their training.

Nevertheless, the correlations are still low. In order to improve the relationship between the ratings and the scores, it is necessary to devise a technique which adds all of the scores in a fashion which duplicates the way that the raters added the factors they used in coming to their decisions. The statistical technique which does this is called multiple correlation.[47]

TABLE 6.--Intercorrelations of Total Test Scores with the Rating Factor Overall Competence by Year of Training

	First and Second N = 109	Third and Fourth N = 119
Diagnostic Interview	.00	.16
Proposed Treatment Interview	.12	.20
Adult Oral	.09	.28*
Multiple Choice	.20	.26*
PMP	-.02	-.01

* Significant at .05

[47]Guilford, Fundamentals, pp. 390-433.

This technique develops an equation which predicts one variable by means of an equation using each of the other variables as elements. Each element has an optional weight determined by the mathematical techniques used to develop the equation.

An example may make this technique more clear. Assume that someone wanted to know the heights of some individuals and all they knew was their weights. They could find out the relationship between weight and height and develop a predictive equation: bw + K = h. The b and K are constants which would help to change the weight figures into inch figures. Now the predicted h's would correlate with the true h's at approximately .55. This is the approximate accuracy of prediction that one can attain using just weight to predict height. Suppose, however, that one was able to obtain some information on waistline. The equation could be improved in this fashion: bw _ cl + K + h. Note especially that waistline has a negative weight because the smaller the waistline the larger the height of individuals of a given constant weight. It often happens in multiple correlation analysis that variables have negative weights when combined with other variables; even when taken alone they have a positive relationship. Thus, waistline, which by itself has a positive relationship with height when combined with weight, has a negative weight in predicting height.

The multiple correlations of the tests and subjects of the various evaluation techniques with the nine rating factors are summarized in Table 7. Note that the relationships have been dramatically increased all through the Table. This means that while each of the techniques has only

TABLE 7.--Correlations and Multiple Correlations of Test Variables Predicting Various Rating Factors

Rating Factor	Multiple R				Highest Single Correlation		
	Reliability of Rating Factor for Total	First and Second Year .109	Third and Fourth Year 119	Total 228	First and Second Year 109	Third and Fourth Year 119	Total 228
Recall	.49	.53*	.53	.43**	.25	.31	.31*
Problem Solving	.69	.51	.56*	.45**	.31	.35	.28*
Information Gathering	.63	.42	.63**	.46**	.30	.44	.33**
Clinical Judgment	.69	.45	.46	.41*	.28	.28	.26*
Patient Relationship	.54	.45	.49	.33	.29	.23	-.17
Colleague Relationship	.67	.51	.53	.34	.29	.32	.19
Surgical Skill	.52	.36	.43	.36	.22	.25	.20
Ethics	.66	.55*	.46	.33	.27	.29	.24
Overall	.72	.48	.51	.37*	.28	.29	.26*

*Significant at .05
**Significant at .01

a small relationship to the rating factors, added together they form a substantial relationship.

The picture becomes even more clear when each of the correlations is squared which gives the percentage of variance in common between the two measures. This information is presented in Table 8. The reliability of the rating factor puts an upper limit on the possible relationship. Note that only 21% of variance in recall remains unexplained, and only 23% of the variance in information gathering. (This is for the third and fourth year residents.) In view of the lack of reliability of some of the tests, these results would seem to indicate that the tests as a group are successfully identifying most of the factors that raters use to decide upon competence. The tests are not just measuring "test wisdom" but traits which have important consequences in other activities of the residents.[48]

It is of particular interest to review the relationships between the variables used to predict the rating factors. The computer program used to obtain the multiple correlations used most[49] of the test variables to obtain the multiple R. However, a few of the tests account for most of the relationships. Table 9 presents the data for two rating form factors. It might be interesting to discuss briefly the results on these factors.

---

[48]A note of caution should be inserted here. The mathematical technique used to develop multiple R's capitalizes on the characteristics of the sample. These relationships must be checked with other samples, a technique known as cross-validation. Plans are already being made to cross validate these relationships.

[49]A few are eliminated by a control on the computer because they make little or no contribution to the prediction.

TABLE 8.--Percentage of Common Variance for Single Test Variables and Multiple Test Variables with Various Rating Factors

Rating Factors	Reliability of Rating Factors	Percentage of Common Variance					
		Multiple Prediction			Single Prediction		
		First and Second Year	Third and Fourth Year	Total	First and Second Year	Third and Fourth Year	Total
	N	109	119	228	109	119	228
Recall	.49	.28	.28	.18	.06	.10	.10
Problem Solving	.69	.26	.31	.20	.10	.12	.08
Information Gathering	.63	.18	.40	.21	.09	.19	.11
Clinical Judgment	.69	.20	.21	.17	.08	.08	.07
Patient Relationships	.54	.20	.24	.11	.08	.05	.02
Colleague Relationships	.67	.26	.28	.12	.08	.10	.04
Surgical Skills	.52	.13	.18	.13	.05	.06	.04
Ethics	.66	.30	.21	.11	.07	.08	.04
Overall Competence	.72	.23	.26	.14	.08	.08	.06

The partial r figure in Table 9 represents the correlation with the factor if all other factors are held constant. For example, if waistline were correlated with height in the general population, a positive relationship would result, because waistline is correlated with weight and weight with height. But if a population of equal weights were selected, waistline would correlate negatively with height. This negative correlation would be the partial correlation of height with waistline when weight is controlled. The equations for computing multiple correlations mathematically control the other variables to obtain the partial r's.

Note that for first and second year residents, PMP Diagnostic II negative has a negative partial correlation. People who score high on Diagnostic Negative scores are those who avoid asking questions or doing procedures which are harmful to the patient. Those who perform these procedures are probably less well-informed, more inquisitive, and more thorough than others.

It may be that the chief of training tends to disregard the lack of information in residents with only one or two years of training but values the curiosity and thoroughness.

The fact that the two PTI's scores had such high and opposite partial r's engenders some speculation. PTI overall competence probably depends upon the combination of two abilities. One is general problem solving skill, the other is ability to interact effectively with patients. If a group of residents have equal scores in overall competence, those with lower scores in interaction would be the better problem solvers.

TABLE 9.--Subtests Most Strongly Related to Selected Rating Form Factors

Rating Form Factor: Ability to Solve Problems

First and Second Year Residents N = 109 Multiple R .51*		Third and Fourth Year Residents N = 119 Multiple R .56*	
Test Variables	Partial r	Test Variables	Partial r
PMP-Problem II-- Diagnosis Neg.	-.27**	Adult Oral	+.22*
PTI-Overall Competence	+.24*	PMP-Problem I- Diagnosis Neg.	+.20*
PTI-Interaction	-.24*		
Multiple Choice-Trauma	+.20*		

Rating Form Factor: Ability to Gather Information

First and Second Year Residents N = 109 Multiple R .42		Third and Fourth Year Residents N = 119 Multiple R .63**	
Test Variables	Partial r	Test Variables	Partial r
PMP-Problem II Diagnosis Neg.	-.22	DI-Diagnosis	+.21*
PTI-Overall Competence	.20	PMP-Problem I- Treatment Neg.	-.20
		Adult Oral	+.20

* Significant at .05 level.
** Significant at .01 level

NOTE: The test variables are all subtests of the various examination techniques. The PMP negative scores are the sums of scores for contra-indicated procedures.

The remaining data in Table 9 is readily understandable. Particularly interesting, however, is the high partial r for the DI diagnosis score on predicting ability to gather information. Chiefs of training seldom observe the process of data gathering, but they often can observe the product. Apparently, when asked about their good data gatherers, they select their good diagnosticians--or those skilled at defending a diagnosis.

The subtests are short and the samples relatively small in this analysis. Without supporting studies, much interpretation can degenerate into rootless speculation. The relationships found, however, are of sufficient magnitude to indicate that the continuation of such studies may prove very valuable in the insights it can provide on the development of competence in orthopaedic surgery.

IX. CONCLUSIONS

Following are some conclusions that have been derived from this study:

(1) Competence in orthopaedic surgery is a multi-factoral concept and some factors have low relationships with others. The critical incident study first established this fact and the patterns of correlations between various tests and rating factors make it even more clear.

(2) Each of the test variables measures important areas of competence not measured by other tests. This conclusion is buttressed by the data in Table 9 and the data on the intercorrelations of the various techniques.

(3) The reliabilities of the oral instruments as presently constituted need to be improved to use these instruments to their fullest potential. This conclusion results from the findings on reliability as reported in Section VII.

(4) The PMP technique may be sampling areas of competence neglected by orthopaedic training programs. See the discussion of construct validity.

(5) Ratings by themselves suffer from various types of observational biases. Other criteria for competence in orthopaedics should be sought. See the sections on conceptual problems and the discussion of Table 9.

X. AREAS FOR FURTHER RESEARCH

As is the case with most fruitful experimental research, this experiment leaves many questions unanswered. Partly as a result of preliminary report of this study. The American Board of Orthopaedic Surgery has revised its examination to improve the reliability of the oral test scores. For example, three exercises similar to the PTI will be given by two examiners in one-half hour. The DI has been combined with some other problem solving exercises so that each candidate will take an hour and one-half problem solving examination.[50]

The rating form has been considerably revised to make the ratings more precise, and ratings have been solicited from two men each for each of the over 800 candidates who will take the 1968 Orthopaedic Certifying Examination. A lengthy PMP section has also been prepared. It is hoped that the replication of the present study on the 1968 examination with its more reliable instruments will buttress or disprove the conclusions established in this paper.

Furthermore, plans are being made to analyze the training programs to see if relationships can be detected between the characteristics of various programs and achievement on various evaluative techniques. This study may point the way to the improvement of the effectiveness and efficiency of training which after all is the ultimate purpose of the development and validation of evaluation techniques.

---

[50]Gregory, Letter.

BIBLIOGRAPHY

Angoff, William H. "Test Reliability and Test Length, "Psychometricka, 19 (1953), pp. 1-16.

Blum, J. Michael and Fitzpatrick, Robert. Critical Performance Requirements for Orthopaedic Surgery, Part I. Method. Pittsburgh, Pa.: American Institutes for Research, 1965.

Cronbach, Lee J. Essentials of Psychological Testing. New York: Harper Bros., 1960.

Flanagan, John G. "The Critical Incident Technique." Psychological Bulletin, No. 4 (1962), pp. 327-358.

Gregory, Charles F. Letter to Examination Committee, American Board of Orthopaedic Surgery. September, 1967.

Guilford, John R. Fundamental Statistics in Psychology and Education. New York: McGraw Hill, 1956.

Holt, Robert R. and Luborsky, Lester. Personality Patterns of Psychiatrists. New York: Basic Books, 1958.

Lewy, Areh, and McGuire, Christine H. "A Study of Alternative Approaches in Estimating the Reliability of Unconventional." Speech read at the Annual Meeting of the American Educational Research Association, Chicago, February 18, 1966.

Levine, Harold G. "Analysis of the Construct Validity of Two Simulation Techniques." Chicago: Center for the Study of Medical Education, University of Illinois College of Medicine, 1967. (Dittoed Report.)

Levine, Harold G. "Report on the January, 1966, Orthopaedic Certification Examination, " Chicago: Center for the Study of Medical Education, University of Illinois College of Medicine, 1967. (Dittoed Report.)

Levine, Harold G. and McGuire, Christine H. "Role Playing as an Evaluative Technique." Journal of Educational Measurement, (In press, 1968).

Lindquist, E. F., ed. Educational Measurement. Chapter 14, "Validity" by Edward Cureton. Washington, D. C.: American Council on Education, 1951.

McGuire, Christine H. and Babbott, David. "Simulation Technique in the Measurement of Problem Solving Skills." Journal of Educational Measurement, 4, No. 4 (1967), pp. 1-10.

Miller, George E., McGuire, Christine H. and Larson, Carroll B. "The Orthopaedic Training Study, "Bulletin of the American Academy of Orthopaedic Surgeons, XIII (1965), pp. 8-11.

U. S. Department of Health, Education and Welfare. Application for Research Grant, No. CH00081-01. "The Efficient Use of Medical Manpower." Chicago, Ill., 1963.

APPENDIX A

Orthopaedic Training Study
American Board of Orthopaedic Surgery
and
Center for the Study of Medical Education
University of Illinois

Critical Performance Requirements for Orthopaedic Surgeons
(derived from The 1964 Critical Incident Study)

I. Skill in Gathering Clinical Information

A. Eliciting Historical Information

1. Obtaining adequate information from the patient
2. Consulting other physicians
3. Checking other sources

B. Obtaining Information by Physical Examination

1. Performing thorough general examination
2. Performing relevant orthopaedic checks

II. Effectiveness in Using Special Diagnostic Methods

A. Obtaining and Interpreting X-rays

1. Directing or ordering appropriate films
2. Obtaining unusual, additional or repeated films
3. Rendering complete and accurate interpretation

B. Obtaining Additional Information by Other Means

1. Obtaining biopsy specimen
2. Obtaining other laboratory data

III. Competence in Developing a Diagnosis

A. Approaching Diagnosis Objectively

1. Double-checking stated or referral diagnosis
2. Persisting to establish definitive diagnosis
3. Avoiding prejudicial analysis

B. Recognizing Condition

1. Recognizing primary disorder
2. Recognizing underlying or associated problem

IV. Judgment in Deciding on Appropriate Care

A. Adapting Treatment to the Individual Case

1. Initiating suitable treatment for condition
2. Treating with regard to special needs
3. Treating with regard to age and general health
4. Attending to contraindications
5. Applying adequate regimen for multiple disorders
6. Inventing, adopting, applying new techniques

B. Determining Extend and Immediacy of Therapy Needs

1. Choosing wisely between simple and radical approach
2. Delaying therapy until diagnosis better established
3. Testing milder treatment first
4. Undertaking immediate treatment

C. Obtaining Consultation on Proposed Treatment

1. Asking for opinions
2. Incorporating suggestions

V. Judgment and Skill in Implementing Treatment

A. Planning the Operation

1. Reviewing literature, X-rays, other material
2. Planning approach and procedures

B. Making Necessary Preparations for Operating

1. Preparing and checking patient
2. Readying staff, operating room, supplies

C. Performing the Operation

1. Asking for confirmation of involved area
2. Knowing and observing anatomical principles
3. Using correct surgical procedures
4. Demonstrating dexterity or skill
5. Taking proper precautions
6. Attending to details
7. Persisting for maximum result

D. Modifying Operative Plans According to Situation

1. Deviating from preplanned procedures
2. Improvising with implements and materials
3. Terminating operation when danger in continuing

E. Handling Operative Complications

1. Recognizing complications
2. Treating complications promptly and effectively

F. Instituting a Non-Operative Therapy Program

1. Using appropriate methods and devices
2. Applying methods and devices correctly

VI. A. Showing Concern and Consideration

1. Taking personal interest
2. Acting in discreet, tactful, dignified manner
3. Avoiding needless alarm, discomfort, or embarrassment
4. Speaking honestly to patient and family
5. Persuading patient to undertake needed care, or only needed care

B. Relieving Anxiety of Patient and Family

1. Reassuring, supporting or calming
2. Explaining condition, treatment, prognosis or complication

IX. Accepting Responsibilities of a Physician

A. Accepting Responsibility for Welfare of Patient

1. Heeding the call for help
2. Devoting necessary time and effort
3. Meeting commitments
4. Insisting on primacy of patient welfare
5. Delegating responsibilities wisely
6. Adequately supervising residents and other staff

B. Recognizing Professional Capabilities and Limitations

1. Doing only what experience permits
2. Asking for help, advice or consultation
3. Following instructions and advice
4. Showing conviction and decisiveness
5. Accepting responsibility for own errors
6. Referring cases to other orthopaedists and facilities

C. Relating Effectively to Other Medical Persons

1. Supporting the actions of other physicians
2. Maintaining open and honest communication
3. Helping other physicians
4. Relating in discreet, tactful manner
5. Respecting other physician's responsibility to his patient

D. Displaying General Medical Competence

1. Detecting, diagnosing, (treating) non-orthopaedic disorders
2. Obtaining appropriate referrals
3. Preventing infection in hospital patients
4. Effectively keeping and following records

E. Manifesting Teaching, Intellectual and Scholarly Attitudes

1. Lecturing effectively
2. Guiding and supporting less experienced orthopaedists
3. Encouraging and contributing to fruitful discussion
4. Contributing to medical knowledge
5. Developing own medical knowledge and skills

F. Accepting General Responsibilities to Profession and Community

1. Serving the profession
2. Serving the community
3. Maintaining personal and intellectual integrity

The Critical Incident Study was carried out with the assistance of The American Institutes for Research, Pittsburgh, Pennsylvania.

APPENDIX B

ORAL EXAMINATION

AMERICAN BOARD OF ORTHOPAEDIC SURGERY

RATING FORM FOR USE WITH "PATIENT" INTERVIEWS

Candidate's Examination Number: _____

(Cols. 1 - 3)

Examiner's Name: _____

(Cols. 4 - 5)

Date: _____

(Cols. 6 - 7)

Starting Time: _____

(Cols. 8 - 11)

Prepared with the assistance of
CENTER FOR THE STUDY OF MEDICAL EDUCATION
UNIVERSITY OF ILLINOIS, COLLEGE OF MEDICINE

Col. No.

12 - 13

RATING OF DIAGNOSTIC INTERVIEW

Diagnostic Case No. _____

Factor I: **Ability to elicit an adequate amount of pertinent information**

Weight 4

(The candidate should ask most of the indicated questions; other questions should be appropriate to the diagnosis.)

14 - 15

01	02	03	04	05	06	07	08	09	10	11	12
☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
Poor			Adequate			Good			Excellent		

Factor II: **Ability to communicate with the patient**

Weight 1

(Did he use appropriate vocabulary, use concepts familiar to the patient, and allow the patient to narrate parts of the history?)

16 - 17

01	02	03	04	05	06	07	08	09	10	11	12
☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
Poor			Adequate			Good			Excellent		

Factor III: **Efficiency in gathering data**

Weight 1

(Did he ask relevant and necessary questions, and avoid the time waste of exploring remote diagnoses which prevent an adequate examination of the pertinent facts?)

18 - 19

01	02	03	04	05	06	07	08	09	10	11	12
☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
Poor			Adequate			Good			Excellent		

Col. No.	RATING OF DIAGNOSTIC INTERVIEW (Continued)
	Factor IV: **Ability to arrive at a diagnosis and present logical reasons for it** Weight 4 (Did he fail to consider all the pertinent facts he uncovered, make errors in relating or interpreting facts, or make errors in weighing the facts at hand?)
20 - 21	01 02 03 Poor ☐ ☐ ☐ 04 05 06 Adequate ☐ ☐ ☐ 07 08 09 Good ☐ ☐ ☐ 10 11 12 Excellent ☐ ☐ ☐
	Factor V: **Overall evaluation of Diagnostic Interview**
22 - 23	01 02 03 Poor ☐ ☐ ☐ 04 05 06 Adequate ☐ ☐ ☐ 07 08 09 Good ☐ ☐ ☐ 10 11 12 Excellent ☐ ☐ ☐
24	Your role: ☐ 1 "patient" ☐ 2 rater only
25	Comments:
	The Candidate was difficult to evaluate because:
26	☐ He spoke slowly
27	☐ He spoke rapidly
28	☐ He did not speak English well
29	☐ He seemed excessively nervous
30	☐ He seemed confused about the procedure
31	☐ Other _____
32	I did not find the Candidate difficult to evaluate ☐

Col. No.

RATING OF PROPOSED TREATMENT INTERVIEW

33 - 34 **Proposed Treatment Case No._____**

Factor I: Effectiveness of the candidate's statements

Weight 6

(Did he give too little information, oversimplify, indicate undue pessimism or optimism, overwhelm the patient with excessive detail or use inappropriate vocabulary?)

35 - 36

01	02	03	04	05	06	07	08	09	10	11	12
☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
Poor			Adequate			Good			Excellent		

Factor II: Effectiveness of the candidate's manner

Weight 2

(Was the manner in which the physician dealt with the "patient" one which would genuinely convince the patient that the physician is interested in his welfare?)

37 - 38

01	02	03	04	05	06	07	08	09	10	11	12
☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
Poor			Adequate			Good			Excellent		

Factor III: Efficiency of the interview in terms of the interaction between patient and physician

(Did the physician present the required information to the patient in a clear-cut efficient fashion?)

39 - 40

01	02	03	04	05	06	07	08	09	10	11	12
☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
Poor			Adequate			Good			Excellent		

Factor IV: Overall evaluation of the Proposed Treatment Interview

41 - 42

01	02	03	04	05	06	07	08	09	10	11	12
☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
Poor			Adequate			Good			Excellent		

43 Your role: ☐ 1 "patient" ☐ 2 rater only

44 Comments:

2

APPENDIX C

RESIDENT EVALUATION FORM

DO NOT WRITE
IN THIS SPACE

Col.No. | Name of Resident_____

1 - 3 | Identification No. ☐☐☐

4 - 6 | Institution_____Code ☐☐☐

7 - 9 | Name of Rater_____Code ☐☐☐

In filling out this form you are to rank the resident on each factor in terms of all the residents in orthopaedic surgery you have known during your career. You are to indicate your rankings by checking the appropriate box under each factor. In making these evaluations DO NOT take into account the resident's level of training. For example, a second year resident may have the potentiality to display outstanding surgical skills, but many fourth year residents might function AT THE PRESENT time at a higher level. He should be ranked lower than they are ranked on surgical skill. If you believe that you do not have sufficient information on the resident to evaluate a particular factor, check the appropriate box. Please write your name in the space above. All the information collected will be held strictly confidential and will not be used for any purpose other than research purposes.

- -

Factor I: Ability to recall factual information concerning general medicine and orthopaedic surgery

Col. No.

This factor deals with the resident's command of the factual information required of a practicing orthopaedist. Residents who score high are those who have a great deal of pertinent information at their "fingertips." Residents who score low are those who consistently display wide gaps in their knowledge. Residents can score well on this factor and low on Factor II below. They may recall a great deal of information, but have difficulty in integrating the information in solving problems in patient treatment and care.

10 | I do not have sufficient information to judge. 1 ☐

RANKING

11-12 | ☐☐☐ ☐☐☐ ☐☐☐ ☐☐☐

01 02 03 | 04 05 06 | 07 08 09 | 10 11 12

Lowest quarter | Third quarter | Second quarter | Highest quarter

Col. No.

Factor II: Ability to use information to solve problems

This factor deals with the resident's effectiveness in using the information he has collected and recalled in solving problems in treatment and diagnosis.

13 I do not have sufficient information to judge. 1 ☐

RANKING

14-15

☐ ☐	☐ ☐	☐ ☐	☐ ☐	☐ ☐	☐ ☐	☐ ☐	☐ ☐	☐ ☐	☐ ☐	☐ ☐	☐ ☐
01	02	03	04	05	06	07	08	09	10	11	12
Lowest quarter			Third quarter			Second quarter			Highest quarter		

- -

Factor III: Ability to gather clinical information

Col. No.

This factor deals with the resident's effectiveness in gathering clinical information. Is he generally thorough and discriminating, or does he fail to gather important information and in general is haphazard and inefficient in this factor?

16 I do not have sufficient information to judge. 1 ☐

RANKING

17-18

☐ ☐	☐ ☐	☐ ☐	☐ ☐	☐ ☐	☐ ☐	☐ ☐	☐ ☐	☐ ☐	☐ ☐	☐ ☐	☐ ☐
01	02	03	04	05	06	07	08	09	10	11	12
Lowest quarter			Third quarter			Second quarter			Highest quarter		

- -

Factor IV: Judgment in deciding on appropriate treatment and care

Col. No.

This factor deals with the resident's ability to properly weigh the many factors involved in deciding on treatment and care, and to come to sound conclusions.

19 I do not have sufficient information to judge. 1 ☐

RANKING

20-21

☐ ☐	☐ ☐	☐ ☐	☐ ☐	☐ ☐	☐ ☐	☐ ☐	☐ ☐	☐ ☐	☐ ☐	☐ ☐	☐ ☐
01	02	03	04	05	06	07	08	09	10	11	12
Lowest quarter			Third quarter			Second quarter			Highest quarter		

- -

Factor V: Skill in surgical procedures

Col. No.

This factor deals with the resident's manipulative skill in carrying out the procedures required of orthopaedists.

22 I do not have sufficient information to judge. ☐ 1

RANKING

23-24 ☐☐ 01 ☐☐ 02 ☐☐ 03 ☐☐ 04 ☐☐ 05 ☐☐ 06 ☐☐ 07 ☐☐ 08 ☐☐ 09 ☐☐ 10 ☐☐ 11 ☐☐ 12

Lowest quarter — Third quarter — Second quarter — Highest quarter

- -

Factor VI: Relating effectively to patients

Col. No.

This factor deals with the resident's tact, consideration and skill in dealing with patients.

25 I do not have sufficient information to judge. ☐ 1

RANKING

26-27 ☐☐ 01 ☐☐ 02 ☐☐ 03 ☐☐ 04 ☐☐ 05 ☐☐ 06 ☐☐ 07 ☐☐ 08 ☐☐ 09 ☐☐ 10 ☐☐ 11 ☐☐ 12

Lowest quarter — Third quarter — Second quarter — Highest quarter

- -

Factor VII: Relating effectively to colleagues and other medical personnel

Col. No.

This factor deals with how effectively the physician works as a member of a medical team, in asking advice, giving advice and showing tact and consideration.

28 I do not have sufficient information to judge. ☐ 1

RANKING

29-30 ☐☐ 01 ☐☐ 02 ☐☐ 03 ☐☐ 04 ☐☐ 05 ☐☐ 06 ☐☐ 07 ☐☐ 08 ☐☐ 09 ☐☐ 10 ☐☐ 11 ☐☐ 12

Lowest quarter — Third quarter — Second quarter — Highest quarter

- -

Col. No.

Factor VIII: Demonstrating the moral and ethical standards required of a physician

This factor deals with the resident's standards in terms of his concern for patients, his financial dealings, and his contacts with other physicians and society in general.

31 I do not have sufficient information to judge. 1 ☐

RANKING

32-33

☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
01	02	03	04	05	06	07	08	09	10	11	12
Lowest quarter			Third quarter			Second quarter			Highest quarter		

Col. No.

Factor IX: Overall competence as an orthopaedic surgeon

34 I do not have sufficient information to judge. 1 ☐

RANKING

35-36

☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
01	02	03	04	05	06	07	08	09	10	11	12
Lowest quarter			Third quarter			Second quarter			Highest quarter		

37-40 Date completed_____